Tutorial: Linear Regression

The Foundation of Predictive Modeling

ES335 - Machine Learning
IIT Gandhinagar

July 23, 2025

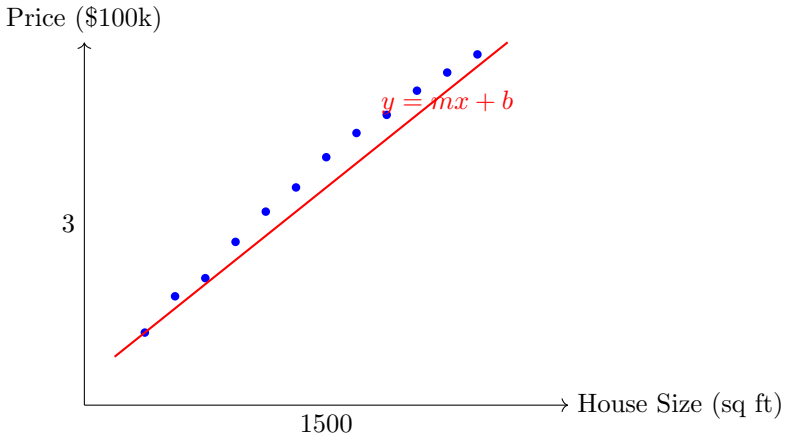
Abstract

Linear regression is the cornerstone of statistical modeling and machine learning. This comprehensive tutorial covers the mathematical foundations, geometric interpretations, optimization methods, and practical applications of linear regression. From simple univariate models to multivariate regression with regularization, learn how to build, evaluate, and interpret linear models effectively.

Contents

1 Introduction: The Prediction Problem

Imagine you're a real estate agent trying to predict house prices. You notice that larger houses tend to cost more. How can you quantify this relationship and make predictions for new houses?



Linear regression finds the **best straight line** through the data that minimizes prediction errors. This simple idea forms the foundation of most machine learning algorithms.

Why Linear Regression?

- **Interpretable**: Coefficients have clear meaning
- **Fast**: Closed-form solution exists
- **Robust**: Well-understood statistical properties
- **Baseline**: Good starting point for any regression problem
- **Foundation**: Basis for many advanced methods

2 Mathematical Foundation

2.1 The Linear Model

For a single feature (simple linear regression):

$$y = \beta_0 + \beta_1 x + \epsilon$$

For multiple features (multiple linear regression):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

In matrix form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where:

- $\mathbf{y} \in \mathbb{R}^n$: target vector
- $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$: design matrix (includes intercept column)
- $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$: parameter vector
- $\boldsymbol{\epsilon} \in \mathbb{R}^n$: error vector

Example #1: Matrix Representation

For 3 houses with features [size, bedrooms]:

$$\mathbf{X} = \begin{bmatrix} 1 & 1500 & 3 \\ 1 & 2000 & 4 \\ 1 & 1200 & 2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 250000 \\ 350000 \\ 200000 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

The model: Price $= \beta_0 + \beta_1 \times \text{Size} + \beta_2 \times \text{Bedrooms}$

2.2 Least Squares Estimation

We find parameters $\boldsymbol{\beta}$ that minimize the **sum of squared errors**:

$$\text{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$$

In matrix form:

$$\text{SSE}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Minimization:

$$\frac{\partial \text{SSE}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

This gives the **normal equations**:

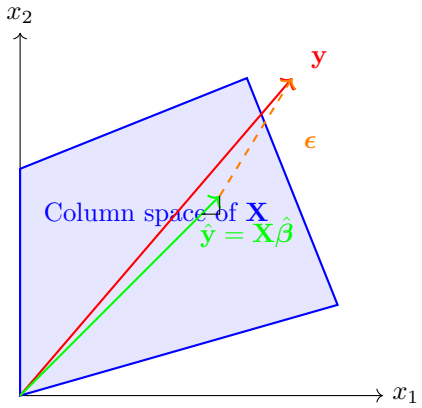
$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$$

Closed-form solution (when $\mathbf{X}^T\mathbf{X}$ is invertible):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

2.3 Geometric Interpretation

Linear regression projects the target vector $\mathbf{y}$ onto the column space of $\mathbf{X}$.



Key insight: $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$. The error vector $\boldsymbol{\epsilon} = \mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to this space.

3 Statistical Properties

3.1 Assumptions

The **Gauss-Markov assumptions** ensure nice statistical properties:

1. **Linearity**: $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$
2. **Independence**: Errors are uncorrelated
3. **Homoscedasticity**: $\text{Var}(\epsilon_i) = \sigma^2$ (constant variance)
4. **No multicollinearity**: $\mathbf{X}$ has full column rank

For inference, we also assume:

5. **Normality**: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

3.2 Properties of the OLS Estimator

Under Gauss-Markov assumptions:
Unbiasedness: $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
Variance: $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
BLUE: Best Linear Unbiased Estimator (minimum variance among all linear unbiased estimators)
Consistency: $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$ as $n \to \infty$

Example #2: Confidence Intervals

For coefficient β_j:

$$\hat{\beta}_j \pm t_{\alpha/2, n-p-1} \times \text{SE}(\hat{\beta}_j)$$

where $\text{SE}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$ and $\hat{\sigma}^2 = \frac{\text{SSE}}{n-p-1}$
Interpretation: We're 95% confident the true coefficient lies in this interval.

4 Model Evaluation

4.1 Goodness of Fit

R-squared (Coefficient of Determination):

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Interpretation:

- $R^2 = 0$: Model explains no variance (as bad as predicting the mean)
- $R^2 = 1$: Model explains all variance (perfect fit)
- $R^2 = 0.7$: Model explains 70% of the variance

Adjusted R-squared (penalizes for additional features):

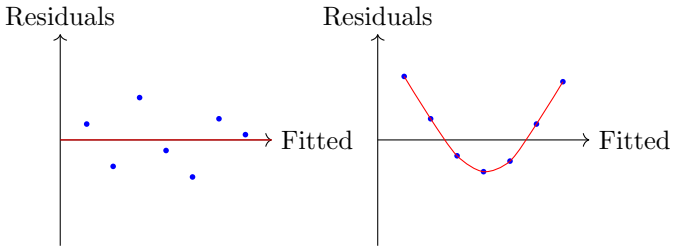
$$R^2_{\text{adj}} = 1 - \frac{(1 - R^2)(n-1)}{n-p-1}$$

4.2 Residual Analysis

Residuals: $e_i = y_i - \hat{y}_i$

Diagnostic plots:

1. **Residuals vs Fitted**: Check linearity and homoscedasticity
2. **Q-Q plot**: Check normality of residuals
3. **Scale-Location**: Check homoscedasticity
4. **Residuals vs Leverage**: Identify influential points



Good: Random scatter **Bad**: Clear pattern

5 Multiple Linear Regression

5.1 Interpretation of Coefficients

In multiple regression, each coefficient represents the **partial effect** of that variable, holding all other variables constant.

$$\frac{\partial y}{\partial x_j} = \beta_j$$

Example #3: House Price Model

$$\text{Price} = 50000 + 100 \times \text{Size} + 15000 \times \text{Bedrooms} + 20000 \times \text{Garage}$$

Interpretation:

- **Baseline**: A house with 0 sq ft, 0 bedrooms, no garage costs \$50,000 (intercept)
- **Size**: Each additional sq ft increases price by \$100, holding bedrooms and garage constant
- **Bedrooms**: Each additional bedroom increases price by \$15,000, holding size and garage constant
- **Garage**: Having a garage increases price by \$20,000, holding size and bedrooms constant

5.2 Multicollinearity

When features are highly correlated, coefficient estimates become unstable.

Detection:

- **Variance Inflation Factor (VIF)**: $\text{VIF}_j = \frac{1}{1-R_j^2}$
- **Condition Number**: $\kappa(\mathbf{X}^T\mathbf{X}) = \frac{\lambda_{\max}}{\lambda_{\min}}$

Rules of thumb:

- VIF > 5: Moderate multicollinearity
- VIF > 10: Severe multicollinearity
- Condition number > 30: Multicollinearity present

Solutions:

- Remove highly correlated features
- Principal Component Regression (PCR)
- Ridge regression (see next section)

6 Advanced Topics

6.1 Weighted Least Squares

When errors have different variances: $\text{Var}(\epsilon_i) = \sigma^2/w_i$

Objective function:

$$\text{WSSE}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$$

Solution:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

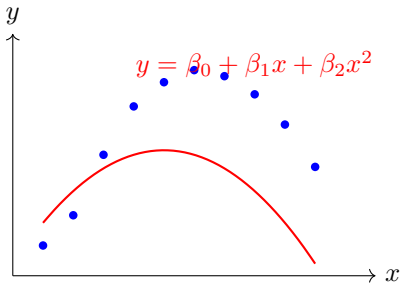
where $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$.

6.2 Polynomial Regression

Extend linear regression to capture non-linear relationships:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d + \epsilon$$

Still linear in parameters! We can use standard linear regression methods.



6.3 Regularization: Ridge Regression

To handle multicollinearity and overfitting, add a penalty term:

$$\text{Ridge}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

Solution:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Effects of λ:

- $\lambda = 0$: Standard OLS
- $\lambda \to \infty$: All coefficients shrink to 0
- Larger λ: More shrinkage, less overfitting

7 Practical Implementation

7.1 Feature Engineering

1. **Scaling**: Standardize features to have mean 0, variance 1

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

2. **Interaction terms**: $x_1 \times x_2$ captures joint effects
3. **Polynomial features**: $x^2, x^3, \ldots$ capture non-linearity
4. **Categorical variables**: One-hot encoding or dummy variables

Example #4: One-Hot Encoding

Original categorical variable: Color $\in$ {Red, Blue, Green}
One-hot encoded:

Original	Color_Red	Color_Blue	Color_Green
Red	1	0	0
Blue	0	1	0
Green	0	0	1

Note: Drop one column to avoid multicollinearity (dummy variable trap)!

7.2 Model Selection

Forward Selection:

1. Start with no features
2. Add feature that most improves model
3. Repeat until no improvement

Backward Elimination:

1. Start with all features
2. Remove feature that least affects model
3. Repeat until significant degradation

Information Criteria:

- **AIC**: $-2 \log L + 2p$
- **BIC**: $-2 \log L + p \log n$

Lower values indicate better models (balance of fit and complexity).

8 Practice Problems

8.1 Basic Problems

Problem #1: Simple Linear Regression

Given data points: (1,2), (2,4), (3,5), (4,7), (5,8)
Find the least squares line $y = \beta_0 + \beta_1 x$.
Solution:

$$\bar{x} = 3, \quad \bar{y} = 5.2$$

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{12}{10} = 1.2$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 5.2 - 1.2(3) = 1.6$$

Line: $y = 1.6 + 1.2x$

Problem #2: Matrix Calculation

For the design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix}$$

Calculate $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
Solution:

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 3 & 9 \\ 9 & 29 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{6}\begin{bmatrix} 29 & -9 \\ -9 & 3 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 21 \\ 71 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

8.2 Intermediate Problems

Problem #3: Multicollinearity Analysis

You have a regression with three predictors. The correlation matrix is:

$$\mathbf{R} = \begin{bmatrix} 1 & 0.2 & 0.95 \\ 0.2 & 1 & 0.3 \\ 0.95 & 0.3 & 1 \end{bmatrix}$$

Calculate VIF for the first predictor.
Solution: Regress x_1 on x_2, x_3:

$$R_1^2 = \text{R-squared from regressing } x_1 \text{ on } x_2, x_3$$

From correlation matrix: $R_1^2 \approx 0.91$ (high correlation with x_3)

$$\text{VIF}_1 = \frac{1}{1 - 0.91} = 11.1$$

This indicates severe multicollinearity!

8.3 Advanced Problems

Problem #4: Ridge Regression Path

For ridge regression with $\lambda \in \{0, 1, 10, 100\}$, describe what happens to: **a)** Coefficient magnitudes **b)** Training error **c)** Test error (assuming overfitting in OLS)
Solution: a) Coefficient magnitudes decrease as λ increases (shrinkage effect) b) Training error increases as λ increases (worse fit to training data) c) Test error first decreases then increases (bias-variance tradeoff)
Optimal λ minimizes test error, balancing underfitting and overfitting.

9 Summary and Best Practices

9.1 When to Use Linear Regression

Ideal scenarios:

- Linear relationship between features and target
- Need interpretable model
- Good baseline for any regression problem
- Sufficient data (more samples than features)
- Features are not highly correlated

Consider alternatives when:

- Strong non-linear relationships
- High-dimensional data (p ¿¿ n)
- Heavy multicollinearity
- Need probabilistic predictions

9.2 Implementation Checklist

- ☐ Check assumptions through residual analysis
- ☐ Handle categorical variables properly (one-hot encoding)
- ☐ Scale features if needed
- ☐ Check for multicollinearity (VIF)
- ☐ Use cross-validation for model selection
- ☐ Consider regularization for high-dimensional data
- ☐ Interpret coefficients in context
- ☐ Validate on held-out test set

9.3 Key Takeaways

1. **Interpretability**: Linear regression provides clear, interpretable relationships
2. **Assumptions matter**: Verify assumptions to ensure valid inference
3. **Feature engineering**: Good features are crucial for performance
4. **Regularization helps**: Ridge/Lasso can improve generalization
5. **Diagnostic tools**: Use residual plots and statistical tests

10 Further Reading

- **Classical Text**: Draper & Smith "Applied Regression Analysis"
- **Statistical Learning**: Hastie et al. "Elements of Statistical Learning"
- **Practical Guide**: James et al. "Introduction to Statistical Learning"
- **Advanced Topics**: Greene "Econometric Analysis"
- **Modern Implementation**: scikit-learn, statsmodels documentation